Automatic Post-Editing Method Using Translation Knowledge Based on Intuitive Common Parts Continuum for Statistical Machine Translation

Hiroshi Echizen'ya[1], Kenji Araki[2], Yuzu Uchida[1], and Eduard Hovy[3]

[1] Hokkai-Gakuen University,
1-1, South-26, West-11, Chuo-ku, Sapporo, 064-0926 Japan
{echi@lst,yuzu@eli}.hokkai-s-u.ac.jp
http://www.lst.hokkai-s-u.ac.jp/~echi/eng-index.html
[2] Hokkaido University,
North-14, West-9, Kita-ku, Sapporo 060-0814 Japan
araki@ist.hokudai.ac.jp
[3] Carnegie Mellon University,
5000 Forbes Avenue, Pittsburgh, PA 15213 USA
hovy@cmu.edu

Abstract. We propose a new post-editing method for statistical machine translation. The method acquires translation rules automatically as translation knowledge from a parallel corpus without depending on linguistic tools. The translation rules, which are acquired based on **I**ntuitive **C**ommon **P**arts **C**ontinuum (ICPC), can deal with the correspondence of the global structure of a source sentence and that of a target sentence without requiring linguistic tools. Moreover, it generates better translation results by application of translation rules to translation results obtained through statistical machine translation. The experimentally obtained results underscore the effectiveness of applying the translation rules for statistical machine translation.

Keywords: Linguistic knowledge, learning method, machine translation, parallel corpus

1 Introduction

For statistical machine translation (SMT), various methods have been proposed. The salient advantage of SMT is that it can process various languages using only a parallel corpus[1,2,3,4]. However, it is difficult for SMT to address the global structure of a sentence because it is based only on the correspondence of local parts, which have adjacent words between the source sentence and the target sentence. To overcome this shortcoming, in SMT, linguistic tools are used in most cases (*i.e.*, POS tagger, parser)[5,6,7]. Those tools are effective for correct analysis of the global structure of a sentence, but it is difficult to translate

A. Ronzhin et al. (Eds.): SPECOM 2014, LNAI 8773, pp. 129–136, 2014.

various languages because few languages have those linguistic tools. Moreover, in previous works of post-editing for MT, various linguistic tools (*i.e.*, a dictionary, parser) have been used[8][9].

Therefore, we propose a new post-editing method for SMT. Our method acquires translation rules as translation knowledge, which can process the global structure of sentence solely from a parallel corpus without the linguistic tools. The translation rules are acquired by recursively determining the common parts between two parallel sentences using determination processes of **I**ntuitive **C**ommon **P**arts **C**ontinuum (ICPC)[10][11]. The parallel sentence represents a pair of a source sentence and target sentence. Moreover, ICPC-based method applies the acquired translation rules to the translation results obtained by SMT. Results show that ICPC-based method, which uses only a parallel corpus, can generate better translation results particularly addressing the global structure of a sentence. Experimentally obtained results using automatic evaluation metrics (*i.e.*, BLEU[12], NIST[13] and APAC[14]) show that the scores produced using ICPC-based method were superior to those obtained using SMT. These results demonstrate the effectiveness of ICPC-based post-editing method.

2 Proposed Method

2.1 Outline

Figure 1 presents the outline of our method. Our method automatically performs post-editing of the translation results obtained using **P**hrase-**B**ased Statistical **M**achine **T**ranslation (PBMT)[3][4]. The PBMT generates the translation model using a parallel corpus. Then it translates the source sentences in the evaluation data. However, in global correspondence between the source sentence and the target sentence, those translation results are insufficient.

Our method acquires translation rules automatically as translation knowledge from a parallel corpus using the determination process of **I**ntuitive **C**ommon **P**arts **C**ontinuum (ICPC). Moreover, the conclusive translation results are generated by combining the translation results obtained using PBMT with the acquired translation rules for the source sentences in the evaluation data. The ICPC-based method is effective at addressing global correspondence between the source sentence and the target sentence using the translation rules.

2.2 Acquisition of Translation Rules Based on ICPC

The translation rules are acquired using common parts between two parallel sentences by the determination process of ICPC. Figure 2 depicts an example of acquisition of translation rule in English-to-Japanese parallel sentences. First, ICPC-based method selects two parallel sentences from the parallel corpus for learning. In Fig. 2, two parallel sentences "(Do you have any liquor or cigarettes ? ; *o sake ka tabako wo o mochi desu ka*[1] *?*)" and "(Do you have any fruits or

[1] Italic indicates the Japanese pronunciation.

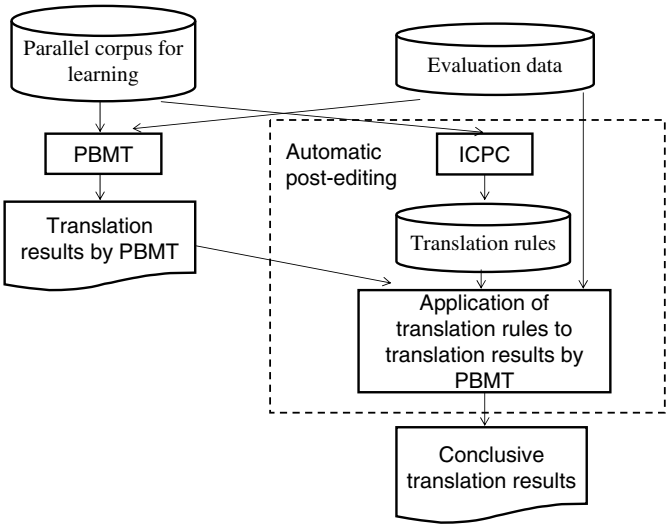

Fig. 1. Outline of our method.

vegetables ? ; *kudamono ka yasai wo o mochi desu ka ?*)" are selected from the parallel corpus. The ICPC-based method determines "Do you have any", "or" and "?" as the common parts of the two English sentences, and "*ka*" and "*wo o mochi desu ka ?*" as the common parts between two Japanese sentences. The different parts are replaced with the variable "*@0*". Consequently, "(Do you have any @0 or @1 ? ; *@0 ka @1 wo o mochi desu ka* ?)" is acquired as the translation rule. This translation rule corresponds to translation knowledge, which indicates the global structure in two parallel sentences. Moreover, it indicates the correspondence between the global structure of the source sentence "Do you have any ... or ... ?" and that of the target sentence "*... ka ... wo o mochi desu ka ?*".

(1) Selection of two parallel sentences

Do you have any liquor or cigarettes ? ; *o sake ka tabako wo o mochi desu ka ?*
Do you have any fruits or vegetables ? ; *kudamono ka yasai wo o mochi desu ka ?*

(2) Extraction of common part

Common parts: (Do you have any @0 or @1 ? ; *@0 ka @1 wo o mochi desu ka ?*)

Fig. 2. Example of translation rule acquisition.

2.3 Application of Translation Rules to Translation Results Obtained Using PBMT

Details of processes using the ICPC-based method are presented below.

(1) The ICPC-based method selects one source sentence from evaluation data and one translation result, which corresponds to the selected source sentence, from the translation results obtained using PBMT.
(2) The ICPC-based method compares the selected source sentence with the source sentence of each translation rule. Then it obtains the translation rules which can fit the source sentence. The translation result obtained using PBMT becomes the conclusive translation result when it cannot obtain any translation rules.
(3) The ICPC-based method calculates similarity based on word-matching between the translation result obtained using PBMT and the target sentence of the selected translation rule. The similarity is less than 1.0. The translation result obtained using PBMT becomes the conclusive translation result when it cannot obtain any translation rule for which the similarity is equal to or greater than threshold 0.4.
(4) The ICPC-based method determines the part, which corresponds to the variable in the target sentence of the translation rule, from the translation result. Moreover, it generates the definitive translation result replacing the variable in the target sentence of the translation rule with the corresponding part in the translation result obtained using PBMT.

Figure 3 depicts an example of generation of the conclusive translation result applying a translation rule. First, ICPC-based method selects "Where is the bus stop for the city center ?" as one source sentence from the evaluation data and "*shinai busu no noriba wa doko kara demasu ka ?*" as the corresponding translation result from the translation results obtained using PBMT. This translation result is the broken Japanese sentence because it corresponds to "Where does the bus stop for the city center leave?" in English. Next, ICPC-based method compares "Where is the bus stop for the city center ?" with "Where is @0 ?", which is the source sentence of the translation rule "(Where is @0 ? ; *@0 wa doko desu ka ?*)". This translation rule can be fit to the source sentence "Where is the bus stop for the city center ?" because "Where is" and "?", which are all parts except variable "@0", are included in the source sentence.

In between the translation result of PBMT "*shinai basu no noriba wa doko kara demasu ka ?*" and the target sentence of the translation rule "*@0 wa doko desu ka ?*", the similarity is 0.4 because the word number of the translation result is 10. Also, the word number of the matching-words is 4. Therefore, the translation rule "(Where is @0 ? ; *@0 wa doko desu ka ?*)" is used as the effective translation rule for generation of the conclusive translation result. The target sentence of translation rule "*@0 wa doko desu ka ?*" possesses the global structure of the correct sentence "*shinai busu no noriba wa doko desu ka ?*".

The ICPC-based method determines the part in the translation result which corresponds to the variable "*@0*" in the target sentence of the translation rule. In

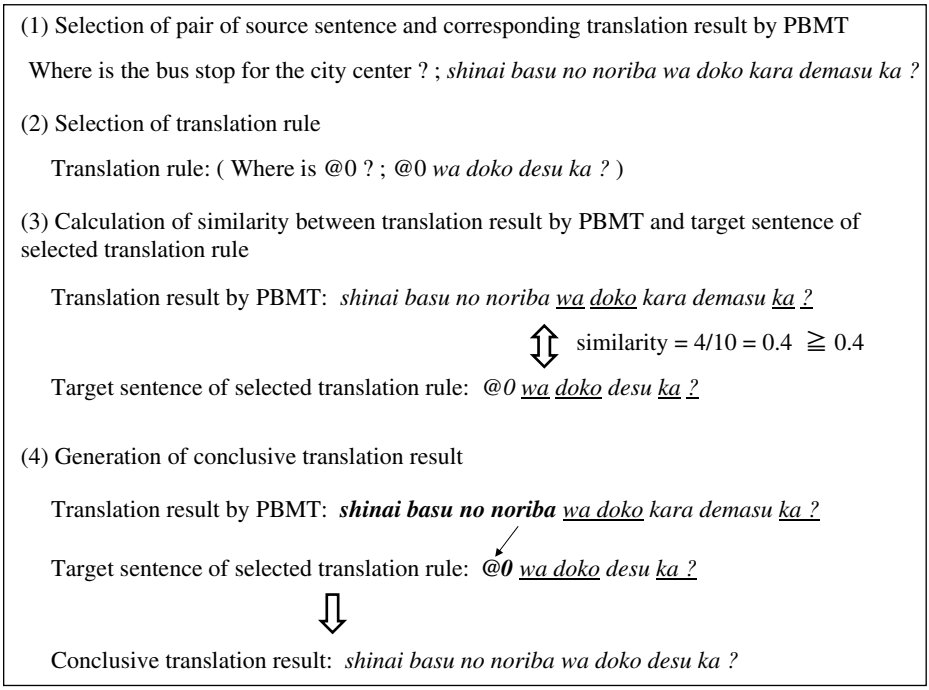

Fig. 3. Example of generation of the conclusive translation result applying a translation rule.

between the translation result "*shinai basu no noriba wa doko kara demasu ka ?*" and the target sentence of the translation rule "*@0 wa doko desu ka ?*", "*shinai basu no noriba*" is determined as the part which corresponds to the variable "*@0*" by the order of appearance of the common parts and different parts. The variable "*@0*" in the target sentence of the translation rule is replaced with "*shinai basu no noriba*". Therefore, "*shinai basu no noriba wa doko desu ka ?*" is obtained as the correct translation result.

The translation rules, which are acquired using the determination process of ICPC, can deal with the global structure of sentence. Namely, they are useful as a framework for both the source sentences and the target sentences. As a result, ICPC-based method can generate these high-quality translation results obtained using only the parallel corpus with no linguistic tools.

3 Experiments

3.1 Experimental Procedure

We used English-to-Japanese travel conversation as the parallel corpus. The number of English-to-Japanese parallel sentences for learning is 1,200. The

English sentences used as the evaluation data are 510. These English-to-Japanese parallel sentences were taken from 10 textbooks for Japanese travelers. The PBMT used 1,200 English-to-Japanese parallel sentences as the learning data, and obtained 510 Japanese translation results for 510 English sentences. In our method, the translation rules were acquired from 1,200 English-to-Japanese parallel sentences using the ICPC determination process. Moreover, ICPC-based method generated 510 Japanese sentences as the conclusive translation results. In PBMT, we used GIZA++[15] as the word alignment model, SRILM[16] as the language model and moses[17] as the translation engine. We compared the translation quality of ICPC-based method with those obtained using PBMT using BLEU, NIST, and APAC as the automatic evaluation metrics. In this case, these metrics use one reference. The APAC indicated high correlation with human judgment among some metrics in WMT2014 when translating from English[14].

3.2 Experimental Results

Table 1 presents scores in the automatic evaluation metrics of ICPC-based method and PBMT. Table 2 exhibits examples of the translation results obtained using ICPC-based method and those from PBMT.

Table 1. Scores for the automatic evaluation metrics.

method	BLEU	NIST	APAC
PBMT	0.0646	1.3832	0.2539
ICPC-based method	0.0635	1.3908	0.2546

Table 2. Examples of translation results.

source sentence	translation result	
	PBMT	ICPC-based method
Is this Mr. Brown ?	*kore wa* Mr. Brown *ka ?*	*kore wa* Mr. Brown *desu ka ?*
Is this Ms. Brown ?	*kore wa* Ms. Brown *ka ?*	*kore wa* Ms. Brown *desu ka ?*
May I speak in English ?	*o namae de* English *mo ii desu ka ?*	*o namae de* English *te mo ii desu ka ?*

3.3 Discussion

Table 1 shows that the scores of NIST and APAC in ICPC-based method are higher than those in PBMT. The BLEU score in the ICPC-based method is lower than that only in PBMT. The reason is that BLEU might be insufficient in two languages for which the structure of source sentence is grammatically different from that of the target sentence[18].

In Table 2, the translation results of PBMT alone are insufficient in the source sentences "Is this Mr. Brown ?" and "Is this Ms. Brown ?" because "*desu*" is

not included in the Japanese translation results. The Japanese word "*desu*" is an extremely important word in this translation result. It corresponds to "is" in English. The translation results of ICPC-based method "*kore wa* Mr. Brown *desu ka ?*" and "*kore wa* Ms. Brown *desu ka ?*" are almost correct. However, "*kore wa*", which corresponds to "this" in English, should be removed from Japanese sentences in telephone conversation scenarios.

Moreover, regarding the source sentence "May I speak in English ?", both the translation result only of PBMT and ICPC-based method are insufficient because "*o namae de* English" is broken as Japanese. However, the translation result of ICPC-based method "*o namae de* English *te mo ii desu ka ?*" is better than that of PBMT "*o namae de* English *mo ii desu ka ?*" because "*te*" is included in the translation result of the ICPC-based method: "*te mo ii desu ka ?*" corresponds to "May I" in English. The ICPC-based method can generate the translation result "*o namae de* English *te mo ii desu ka ?*" using the translation rule "(May I @0 ? ; *@0 te mo ii desu ka ?*)". This translation rule is useful as a frame for the source sentence "May I speak in English ?". Therefore, ICPC-based method produced better translation results "*o namae de* English *te mo ii desu ka ?*" using the translation rule "(May I @0 ? ; *@0 te mo ii desu ka ?*)", which can accommodate the global structure of sentence.

The translation results that were improved using the translation rules were 17 among all 510 translation results. The number of the effective translation rules acquired by the determination process of ICPC was insufficient, although about 2,000 translation rules were acquired. The ICPC-based method must acquire the translation rules efficiently from a parallel corpus. For example, it is effective to use the statistical information when acquiring effective translation rules. The scores of the evaluation metrics improve by increasing the effective translation rules in ICPC-based method.

4 Conclusion

As described herein, we propose a new post-editing method that uses translation rules acquired by the ICPC determination process. The ICPC-based method can process the global structure of sentences using the acquired translation rules only from a parallel corpus with no linguistic tools. Therefore, ICPC-based method is effective for the various languages. Future studies are expected to improve ICPC-based method to acquire more translation rules using statistical information, and to perform the evaluation results using the various languages.

References

1. Brown, P.F., Cocke, J., Pietra, S.A.D., Pietra, V.J.D., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roosin, P.S.: A Statistical Approach to Machine Translation. Computational Linguistics 16(2), 79–85 (1990)
2. Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L.: The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics 19(2), 263–311 (1993)

3. Koehn, P., Och, F.J., Marcu, D.: Statistical Phrase-based Translation. In: Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pp. 48–54 (2003)
4. Chiang, D.: A Hierarchical Phrase-based Model for Statistical Machine Translation. In: Proc. of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 263–270 (2005)
5. McDonald, R., Crammer, K., Pereira, F.: Online Large-Margin Training of Dependency Parsers. In: Proc. of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 91–98 (2005)
6. Chiang, D., Marton, Y., Resnik, P.: Online Large-Margin Training of Syntactic and Structural Translation Features. In: Proc. of the Conference on Empirical Methods in Natural Language Processing, pp. 224–233 (2008)
7. Cherry, C., Moore, R.C., Quirk, C.: On Hierarchical Re-ordering and Permutation Parsing for Phrase-based Decoding. In: Proc. of the Seventh Workshop on Statistical Machine Translation, pp. 200–209 (2012)
8. Dugast, L., Senellart, J., Koehn, P.: Statistical Post-Editing on SYSTRAN's Rule-Based Translation System. In: Proc. of the Second Workshop on Statistical Machine Translation, pp. 220–223 (2007)
9. Plitt, M., Masselot, F.: A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. The Prague Bulletin of Mathematical Linguistics 93, 7–16 (2010)
10. Echizen-ya, H., Araki, K.: Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum. In: Proc. of the Eleventh Machine Translation Summit, pp. 151–158 (2007)
11. Echizen'ya, H., Araki, K., Hovy, E.: Optimization for Efficient Determination of Chunk in Automatic Evaluation for Machine Translation. In: Proc. of the 1st International Workshop on Optimization Techniques for Human Language Technology (OPTHLT 2012) / COLING 2012, pp. 17–30 (2012)
12. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
13. Doddington, G.: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In: Proc. of the second International Conference on Human Language Technology Research, pp. 138–145 (2002)
14. Echizen'ya, H., Araki, K., Hovy, E.: Application of Prize based on Sentence Length in Chunk-based Automatic Evaluation of Machine Translation. In: Proc. of the Ninth Workshop on Statistical Machine Translation, pp. 381–386 (2014)
15. Och, F.J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics 29(1), 19–51 (2003)
16. Stolcke, A.: SRILM – An Extensible Language Modeling Toolkit. In: Proc. of the International Conference on Spoken Language Processing, pp. 901–904 (2002)
17. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proc. of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 177–180 (2007)
18. Isozaki, H., Sudoh, K., Tsukada, H., Duh, K.: Head Finalization: A Simple Reordering Rule for SOV Languages. In: Proc. of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pp. 244–251 (2010)